The School of Information and its Relationship to Computer Science at UC Berkeley

Zachary A. Pardos, Andrew Joo Hun Nam
University of California, Berkeley

Abstract
There is a polymorphism to Schools of Information that allows for a nimbleness of research that can travel outside of academically normative paths. The fuzzy borders between the school and its nearest departments are a product of the interdisciplinarity of its faculty composition and nature of the problems it takes on. The boundaries of the school are brought into relief as data science curriculum and research grow and the relationship of the school to computer science is necessarily considered. In this paper, we investigate the iSchool's relationship to Computer Science through the lens of course enrollment behavior of over 160,000 students at UC Berkeley between 2007 and 2015.
Keywords: iSchools; Computer Science; Course enrollment; Descriptive statistics; Distributed representation
Citation: Pardos, Z. A., & Nam, A. J. H. (2017). The School of Information and its Relationship to Computer Science at UC Berkeley. In *iConference 2017 Proceedings* (pp. 309–316). https://doi.org/10.9776/17030

Acknowledgements: We would like to thank the following UC Berkeley staff for their assistance in procuring and describing these data: Mark Chiang, Andrew Eppig, Oliver Heyer, Max Michel, Aswan Movva, Jenn Stringer, and Walter Wong.
Contact: zp@ischool.berkeley.edu

1 Introduction

As research and teaching surrounding data science grows at UC Berkeley, and other institutions with schools of Information, the inevitable question arises of the iSchool's relation to Computer Science and its potential overlap at those institutions. In this paper, we investigate this relationship through the lens of descriptive analyses of the student majors served by the two departments[1] and through a novel method comparing their similarity in vector space using a distributed representation of courses.

2 Related Work

The Institutional self-study of Schools of Information has been a common practice (Larsen, 2009; Olson & Grudin, 2009) due to the iSchools' multitude of roots, some born out of Library Information Science (LIS), some from business and management, and others out of Computing. Wiggins & Sawyer (2012) studied the schools from the perspective of faculty composition and recruitment. Their clustering put UC Berkeley's iSchool in the category of "Sociotechnical," suggesting a 24-40% computing composition with 3+ additional representatives from the social sciences and no more than 34% from LIS. Wu et al. (2012) further investigates the faculty of iSchools, finding among the 25 iSchools at the time, 28% of faculty had a degree background in LIS and an equal percentage had a degree background in CS. The second tier of most common backgrounds consisted of business, engineering, and education. Wu et al. categorized the Berkeley iSchool as a Business and Management (BM) iSchool. In their study, BM iSchools most commonly were hosts to master's degrees in information management and systems, while CS iSchool were more commonly host to degrees in CS and human-computer interaction. Analyzing the research interests, research funding, and degree programs offered by various iSchools, Wu et al. conclude the iSchools' foci are clearly centered on HCI, intelligent systems, information theory, and network technology, yet still acknowledge the

[1] The organization of campus units at UC Berkeley designates Computer Science as a subject area of the department of EECS. The School of Information is not a department but rather a professional school. For the sake of simplicity, both will be referred to as "departments" in this paper.

interdisciplinary identity of the information field. Thompson (2008) breaks down the breadth of iSchools to be related primarily to philosophy and sociology at the high levels of curricular abstraction, with lower level foundational principles most closely shared with computer science and electrical engineering. Given these foundational and epistomological overlaps with CS, it is no surprise that the topic of iSchools relationship to CS is of frequent and ongoing study. Faculty hiring reflects the direction of a unit and faculty teaching can be considered a loose proxy for their research reflecting their training. The variety of students attracted to these courses are a further reflection of the school and how it is perceived by students in other disciplines. We investigate the UC Berkeley iSchool from this perspective of course enrollment.

3 Dataset

We used anonymized enrollment information of graduate and undergraduate students attending the University of California, Berkeley between Fall 2008 and Fall 2015. The dataset contained course IDs and declared major(s) for every enrolled semester for 99,971 UC Berkeley undergraduates, 38,466 graduates, and 22,814 visiting summer exchange students. Students declared 259 unique degree programs hosted by 208 academic departments offering a total of 9,739 primary lecture courses. The major associated with an enrollment is the student's declared major at the time of that enrollment.

4 Methods

4.1 Representing Courses and Departments as Vectors

Using the course enrollment data of each student, we encoded each semester as a multi-hot vector, where each unique course would be treated as a 1 and all other courses as a 0. Using the skip-gram prediction model (Figure 1), we trained an artificial neural network to predict other courses the student had taken using a context of the previous and next N courses to the input course. This N is a hyper parameter of the skip-gram called window size. This prediction scheme encodes the courses into a vector represented by the set of weights learned associated with the course's one-hot and the intermediary hidden layer. We used the resulting vector representation of the courses learned by the neural network to compare departments to one another by taking the average vector of the courses within each department and calculating the cosine distance between each other. While the mechanism for training the network involves predicting courses in context, it should be noted that it is the resultant course vector that is the desired element of the trained model. This scheme of predicting context is a variant on traditional auto-encoder approaches (Hinton & Salakhutdinov, 2006) but with a similar goal of modeling the structure of the input data.

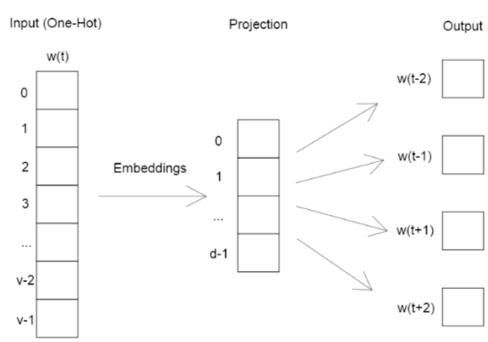

Figure 1. Skip-Gram Model

As a pre-preprocessing step, departments with fewer than 10 courses were removed as were courses with fewer than 10 total students enrolled over the study period.

The methods for analyzing departments used the unit vector average of course vectors, enabling cosine similarity (Equation 1) of the vectors to be used to evaluate the relative closeness between one department and another, where **a** and **b** are department vector representations.

$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \mathbf{b}}{\|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2}$$

Equation 1. Cosine Similarity

To measure the validity of our approach, we trained a simple feed-forward neural network using the vector representations as inputs to predict the corresponding departments of each of the courses. We used Tree-structured Parzen Estimator (TPE) algorithm by Bergstra et al. (2010;2013) to optimize the topology and other hyperparameters of the neural network. The best topology consisted of a single 218-neuron layer with PReLU activation and a dropout rate of .30997, optimized using Adagrad with a learning rate of .02879. We used the softmax function as the output activation, a batch size of 100, and an early stopping with patience of 2. Categorical cross-entropy (Equation 2) was used as the loss function which is equal to the negative log of the probability of the correct label, *q(x)*. The *p(x)* term is 1 for the correct label and 0 for all other labels. The neural network predicted the department of test set of course vectors with 79.4% accuracy using the best model, far above the 7.07% baseline of guessing the department with the highest number of courses (Law had 639 out of 9038 courses used in our model). This result is significant because the vectors themselves were trained in a relatively unsupervised fashion, not optimized for the particular objective of identifying the department. This gave us confidence that the vector representations of courses contained sufficient richness to be used for the task of mining the relationships between the course vector representations.

$$\mathrm{H(p,q)} = -\sum_{x} p(x) \log q(x)$$

Equation 2. Categorical Cross-Entropy

5 Results

In this section, we present results on the comparison of the two departments using a descriptive statistics analysis of the majors attracted to each (5.1). This is followed by an analysis of which departments are most similar to the iSchool and CS (5.2) and the composition of both schools expressed as a sum of two other departments (5.3). The results of the analyses in sections 5.2 and 5.3 utilized the skip-gram model.

5.1 Major Distribution of Students Who Take iSchool and Computer Science Courses

The different student majors attracted to courses offered by a particular department is a reflection of the disciplines with which that department resonates with. It similarly reflects students' perceptions of what departments are relevant to their interests from the standpoint of their primary area of study. By looking at the makeup of outside-of-department majors enrolled in graduate courses in the iSchool and CS, we can observe how the two schools differ in the communities they serve.

Rank	Major	All Students		Undergraduates		Graduates	
1	Electrical Engineering and Computer Science	610	(8%)	308	(43%)	302	(4%)
2	Business Administration$_{10}$	162	(2%)	34	(5%)	128	(2%)
3	Public Policy	86	(1%)	0	(0%)	86	(1%)
4	Mechanical Engineering$_{1}$	81	(1%)	7	(1%)	74	(1%)

5	Public Health	63	(1%)	0	(0%)	63	(1%)
6	Interdisciplinary Studies	54	(1%)	54	(8%)	0	(0%)
7	Education	50	(1%)	0	(0%)	50	(1%)
8	Industrial Engineering and Operations Research$_6$	49	(1%)	2	(0%)	47	(1%)
9	Architecture	46	(1%)	2	(0%)	44	(1%)
10	Cognitive Science$_8$	42	(1%)	42	(6%)	0	(0%)
	Total (Outside iSchool)	1,834	(24%)	709	(100%)	1,125	(16%)
	Total (Inside iSchool)	5,953	(76%)	0	(0%)	5,953	(84%)
	Total (Overall)	7,787	(100%)	709	(100%)	7,078	(100%)

Table 1. Number and proportion of students who enrolled in graduate level iSchool classes by major[2,3]

Rank	Major	All Students		Undergraduates		Graduates	
1	Mechanical Engineering$_6$	274	(2%)	6	(0%)	268	(3%)
2	Information	215	(2%)	0	(0%)	215	(2%)
3	Mathematics	135	(1%)	67	(3%)	68	(1%)
4	Statistics	122	(1%)	26	(1%)	96	(1%)
5	Applied Mathematics	121	(1%)	96	(4%)	25	(0%)
6	Industrial Engineering and Operations Research$_8$	102	(1%)	0	(0%)	102	(1%)
7	Civil and Environmental Engineering	97	(1%)	0	(0%)	97	(1%)
8	Cognitive Science$_{10}$	91	(1%)	91	(4%)	0	(0%)
9	Physics	64	(1%)	12	(1%)	52	(1%)
10	Business Administration$_2$	52	(0%)	29	(1%)	23	(0%)
29	Education	11	(0%)	0	(0%)	11	(0%)
	Total (Outside EECS)	1,858	(15%)	516	(23%)	1,342	(14%)
	Total (Inside EECS)	10,300	(85%)	1,762	(77%)	8,538	(86%)
	Total (Overall)	12,158	(100%)	2,278	(100%)	9,880	(100%)

Table 2. Number and proportion of students who enrolled in graduate level Computer Science classes by major[4]

Worthy of note is the high percentage of the outside community the iSchool serves; 24% of all of its enrolled students are from an outside major vs. EECS (graduate courses) serving 15% outside of their department. Among the top 10 majors drawn to iSchool courses, 37% of the total students are majors from the humanities and social sciences compared to Computer Science serving 4% from those disciplines.

As seen in Table 1, of the 7,787 student enrollments in iSchool classes in our dataset, only 76% of them were students from the Masters in Information Management Science program or the Masters in Data Science program. Computer Science and Business Administration, two majors often associated with iSchool, were the top 2 outside majors which composed iSchool classrooms, with 8% and 2% of the population respectively. The remaining 14% that came from other disciplines indicate a diverse set of academic fields that the iSchool attracts, ranging from highly technical fields such as Mechanical Engineering to social

[2] Subscripted numbers next to department names in Table 1 indicate rank in Table 2 and vice-versa.

[3] STEM majors are underlined (as defined by the U.S. Department of Homeland Security: https://www.ice.gov/sites/default/files/documents/Document/2016/stem-list.pdf)

[4] Information represents Berkeley's Master's in Information Management Systems program

science such as Media Studies and Political Science. Of the 1,834 enrollments in iSchool courses from students outside of iSchool's degree programs, 39% of them were undergraduate students.

Table 2 shows that Computer Science mainly attracted students studying EECS (85%) and other highly technical fields such as Mechanical Engineering and Mathematics. iSchool students were the 2nd largest audience in Computer Science classrooms after Mechanical Engineering students. Business Administration followed at a distant 10th, suggesting that while they may be related, the relationship is not as strong as it is with the iSchool.

Education and Information have common roots in their shared human-centered study of the generation, curation, and delivery of information and knowledge. This relationship is neither confirmed nor rejected with the moderate level of cross-enrollments observed at Berkeley. Education students ranked 7th in the composition of iSchool classes, but this relationship was not reciprocal, with iSchool students ranking 36th in the composition of Education classes.

5.2 Closest Departments to the iSchool and Computer Science in Cosine Distance

In this study, we used a unit average of a department's course vectors to represent the vector for that department and found the departments closest in cosine distance to the iSchool and Computer Science.

The top results for nearest department vectors were New Media and Business Administration (MBA), with Computer Science coming in 5th. Not listed in the table was Design Innovation, ranked 2nd before being filtered out for having fewer than 10 courses. While in the study performed by Wu et al. (2012) on iSchools, topics such as user modeling and HCI were listed under Computer Science schools, Berkeley's institutes for New Media and Design Innovation may have specialized further to address these topics. Among the 10 nearest departments to the Information vector, 5 could be classified as highly technical, such as Computer Science and Vision Science, while others were social sciences and humanities. Interestingly, the cosine similarity between MBA and Computer are relatively high, suggesting that Berkeley's iSchool is in near-equal balance between Business Administration and Computer Science.

Electrical Engineering is the nearest neighbor to Computer Science with a cosine similarity of .649 with Information as a distant second with a similarity of .491. Among the 10 nearest departments to Computer Science, 7 of them are STEM subjects such as Mathematics, Statistics, and Engineering. Among the top 10 nearest neighbors, the iSchool and CS have four mutual neighbors; Electrical Engineering, Industrial Engineering and Operations Research (IEOR), Mechanical Engineering, and New Media.

Rank	Department	Similarity
1	New Media$_7$	0.610
2	Business Administration (MBA)	0.541
3	Development Practice	0.518
4	Electrical Engineering$_1$	0.494
5	Computer Science	0.491
6	Industrial Engineering and Operations Research$_4$	0.488
7	Public Policy	0.439
8	Vision Science	0.425
9	Education	0.410
10	Mechanical Engineering$_9$	0.403

Table 3. Nearest Department to Information by Cosine Similarity

Rank	Department	Similarity
1	Electrical Engineering$_4$	0.649
2	Information	0.491
3	Mathematics	0.489
4	Industrial Engineering and Operations Research$_6$	0.485
5	Engineering	0.475
6	Statistics	0.465
7	New Media$_1$	0.420
8	Cognitive Science	0.386
9	Physics	0.382
10	Mechanical Engineering$_{10}$	0.325
77	Education	-0.011

Table 4. Nearest Department to Computer Science by Cosine Similarity

5.3 Department Composition as a Sum of Other Departments

In this study, we investigate which two departments best comprise the iSchool and CS by looking for the two department vectors, that when added together, are closest to those departments.

Among the top vector additions that equate to Information is the sum between Business Administration and Computer Science. The frequency of both social sciences and engineering departments in the results suggest that the iSchool offers an intersection between the social and technical sciences. Computer Science and New Media appeared in 5 of the top 10 vector equations for Information, while Business Administration appeared twice.

For Computer Science, Electrical Engineering appeared in all of the top 10 vector additions, likely due to Berkeley's combination of Computer Science and Electrical Engineering subject areas into the department of EECS. For a more nuanced analysis, we filtered out additions involving Electrical Engineering. Within this set, the most common departments were Mathematics and IEOR with 4 appearances each. Information appeared in 3 equations in combination with Mathematics, Statistics, and IEOR.

Rank	Department 1	Department 2	Cosine Similarity
1	Development Practice	New Media	0.721
2	Business Administration (MBA)	Computer Science	0.710
3	Computer Science	Development Practice	0.706
4	Business Administration (MBA)	New Media	0.678
5	Computer Science	Public Policy	0.670
6	New Media	Public Policy	0.665
7	Electrical Engineering	New Media	0.656
8	Industrial Engineering and Operations Research	New Media	0.650
9	Computer Science	Journalism	0.646
10	Computer Science	Education	0.640

Table 5. Department 1 + Department 2 similarity to Information

Rank	Department 1	Department 2	Cosine Similarity
1	Information	Mathematics	0.633
2	Mathematics	New Media	0.614
3	Cognitive Science	Industrial Engineering and Operations Research	0.602
4	New Media	Statistics	0.597
5	Mathematics	Industrial Engineering and Operations Research	0.593
6	Linguistics	Industrial Engineering and Operations Research	0.587
7	Information	Statistics	0.580
8	Information	Industrial Engineering and Operations Research	0.566
9	New Media	Industrial Engineering and Operations Research	0.565
10	Cognitive Science	Mathematics	0.562

Table 6. Department 1 + Department 2 similarity to Computer Science (EE excluded)

6 Discussion

6.1 Composition of iSchool and CS Courses

Our analysis of the students that compose iSchool as well as the arrangements of the department vectors shows that iSchool is a diverse program that draws students from Computer Science, Business Administration, and various social sciences. The students' course selections reflected in descriptive statistics of iSchool classrooms reinforce the association between Information, Computer Science, and Business Administration. The Computer Science and MBA departments' vectors similarity to the Information department's vector suggests that the two are viewed with roughly equal relevance to Information by students. Comparing the students in iSchool and Computer Science courses, the iSchool is more diverse in its enrollees with respect to STEM and non-STEM backgrounds, while the Computer Science classroom mainly consisted of students from STEM majors. Moreover, iSchool attracted more undergraduates from outside its department than did Computer Science. The vector representations revealed a similar result, associating several social sciences with the iSchool vector while mainly associating Computer Science with other technical departments.

6.2 Limitations

The dataset used in this study restricted the analysis on the composition and focus of iSchools to strictly that of UC Berkeley and may not be adequately representative of iSchools across other campuses. Moreover, UC Berkeley's transition towards adding more design oriented courses through the new Jacobs Institute for Design Innovation is relatively young and was mostly excluded from out dataset due to insufficient course offerings to date.

7 Conclusion

In this study, we examined the composition of UC Berkeley's graduate level Information program in comparison to its graduate level Computer Science program. Specifically, we can draw the following conclusions in our comparisons of the two units.

1. Among the iSchool's outside major students, more than 1/3rd were from the social sciences and humanities (STEM) vs. CS where less than 1/20th were from those disciplines.
2. EECS is the top outside major served by iSchool courses (with half of their students having been undergraduates)
3. Information is Computer Science's 2nd most served outside major (very close to the top outside major, MechEng)

4. The iSchool has served 2/3rd the number of students Computer Science has served (graduate courses only)
5. Information and Education appear to have some relationship, but are distant compared to the iSchool's presence in Business Administration and Computer Science, reinforcing the Wu et al. (2012) categorization of the Berkeley iSchool as in the Business Management category.

7.1 Future Work

Using the vector representations of courses and departments, we plan to explore the diversity of iSchool courses, examining the spread of the course vectors in comparison to other departments as well as mapping iSchool and CS courses to a vector projection representing Data Science.

8 References

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems* (pp. 2546-2554).

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *ICML (1), 28*, 115-123.

Dillon, A. (2012). What it means to be an iSchool. *Journal of Education for Library and Information Science*, 267-273.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504-507.

Larsen, R. L. (2009). iSchool. *Encyclopedia of Library and Information Science*, Third edition, Taylor and Francis, 3018-3023.

Mikolov, T., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Olson, G., & Grudin, J. (2009). Timelines, the information school phenomenon. *Interactions*, 16(2), 15-19.

Thompson, R. A. (2008). I-School Curricula: How Wide? How Deep? In *Proceedings of the iConference.* Los Angeles, CA.

Wiggins, A., & Sawyer, S. (2012). Intellectual diversity and the faculty composition of iSchools. *Journal of the American Society for Information Science and Technology*, 63(1), 8-21.

Wu, D., He, D., Jiang, J., Dong, W., & Vo, K. T. (2012). The state of iSchools: an analysis of academic research and graduate education. *Journal of Information Science*, 38(1), 15-36.